IDEAS

Journal on Language Teaching and Learning, Linguistics and Literature

ISSN 2338-4778 (Print)
ISSN 2548-4192 (Online)

Volume 14, Number 2, December 2026
pp. 7052 - 7070

Issued by English study program of UIN Palopo

Improving SMK Students' Speaking Skills with AI-Based Learning in Vocational Contexts: A Quasi-Experimental Study

Sitti Maesura Yahya[1], Munir[2], Amra Ariyani[3]
[1,2,3]Postgraduate Program of English Language Education, State University of Makassar (UNM), Makassar, Indonesia

Article Info

Received: 2026-05-29
Revised: 2026-06-09
Accepted: 2026-09-03

Keywords:
AI-based learning;
Gemini;
speaking skills;
vocational context;
quasi-experimental

DOI:
10.24256/ideas. v14i2. 10887

Corresponding Author:
Munir
munir@unm.ac.id
Postgraduate Program of English Language Education, State University of Makassar (UNM), Makassar, Indonesia

Abstract

English speaking proficiency is essential for vocational high school (SMK) graduates entering the workforce, yet speaking instruction often lacks authentic workplace contexts and scalable feedback mechanisms. Although AI-based language learning and vocational contextualized approaches are each documented in the literature, their synergistic integration remains empirically under-researched in Indonesian SMK pharmacy settings. This study examined whether AI-based learning embedded in vocational pharmacy tasks significantly improves the English speaking skills of Grade XI students at SMKN 1 Polewali. A quasi-experimental nonequivalent pretest–posttest control group design was employed with 31 Grade XI Pharmacy students (experimental n = 15; control n = 16). The experimental group received Gemini-assisted patient-counseling role-play instruction; the control group received conventional vocational-contextualized instruction without AI, across eight sessions over four weeks. Speaking performance was assessed using Heaton's (1988) rubric covering accuracy, fluency, and comprehensibility, and analyzed using paired-samples t-tests, an independent-samples t-test, and Cohen's d. The experimental group demonstrated significant pre–post improvement across all speaking aspects ($p < .001$) with large effect sizes (average $d = 2.97$), and significantly outperformed the control group at posttest ($p = .006$, $d = 1.06$). Comprehensibility showed the greatest gain. Theoretically, the study extends Swain's (1995) Output Hypothesis into AI-assisted ESP contexts by showing how chatbot-mediated role-play supports pushed output and immediate corrective feedback in resource-constrained SMK settings. The novelty lies in the synergistic integration of AI-based learning and vocational contextualized instruction—an approach not previously

validated in Indonesian SMK pharmacy classrooms. AI-based learning in vocational contexts effectively improved speaking skills. Integrating accessible AI tools with workplace-relevant tasks offers a scalable model for English instruction in SMK programmes.

1. Introduction

Rapid digitalisation and cross-border labour mobility in Southeast Asia have contributed to intensified demand for communicative English among vocational high school (SMK) graduates in Indonesia (Thompson & Lim, 2021). English functions as a global lingua franca that supports workplace interaction across diverse occupational fields (Seidlhofer, 2011). For SMK learners who enter employment soon after graduation, spoken English is increasingly treated as a vocational asset rather than an academic credential alone (Safira & Azzahra, 2022). National labour-market indicators and employer surveys suggest persistent skills–employment gaps in sectors requiring professional communication (Safira & Azzahra, 2022; Thompson & Lim, 2021). Within this context, oral English instruction in SMK programmes should aim to prepare learners to perform authentic workplace communication, not merely to complete classroom exercises.

Pharmacy-oriented SMK programmes illustrate this vocational speaking requirement with particular clarity. ESP research identifies patient counseling, medication instruction, and professional service talk as core oral competencies for pharmacy practitioners (Anthony, 2018; Dudley-Evans & St. John, 1998). The present study was situated at SMKN 1 Polewali, West Sulawesi, where Grade XI Pharmacy students encounter speaking constraints widely reported in Indonesian SMK contexts, including limited vocabulary, grammatical inaccuracy, speaking anxiety, and insufficient opportunities for sustained oral practice (Efrizah et al., 2024). Many learners therefore struggle to transfer classroom English into vocationally authentic communication tasks. Addressing this problem may require instructional models that combine meaningful workplace contexts with scalable practice and feedback mechanisms appropriate for resource-constrained SMK environments.

Parallel developments in educational technology have positioned artificial intelligence (AI) as a promising support for speaking development. AI-based learning integrates intelligent systems—often powered by natural language processing—into language learning to provide adaptive interaction, immediate feedback, and repeated practice beyond conventional classroom time (Zawacki-Richter et al., 2019; Lima et al., 2024). For speaking instruction, AI tools can support pronunciation feedback, grammar and vocabulary correction, and conversational simulation (Zou et al., 2023; Nguyen, 2024), and recent studies suggest that AI-assisted activities may improve oral performance in controlled conditions (Safitri et al., 2025; Nhan et al., 2025).

In Indonesian educational settings, reviews of AI adoption indicate that chatbot-mediated interaction may supplement limited classroom practice (Nuryadin & Marlina, 2023); accessible platforms such as Google Gemini offer a low-cost option for extending vocational speaking rehearsal in SMK environments where smartphone access is available. Nevertheless, most AI-speaking research prioritizes general proficiency in mainstream populations rather than vocationally situated ESP tasks in authentic workplace scenarios (Nguyen, 2024; Nhan et al., 2025). How AI-mediated feedback functions alongside ESP content and task-based speaking cycles therefore requires clearer theoretical and empirical articulation.

ISSN 2338-4778 (Print)
ISSN 2548-4192 (Online)

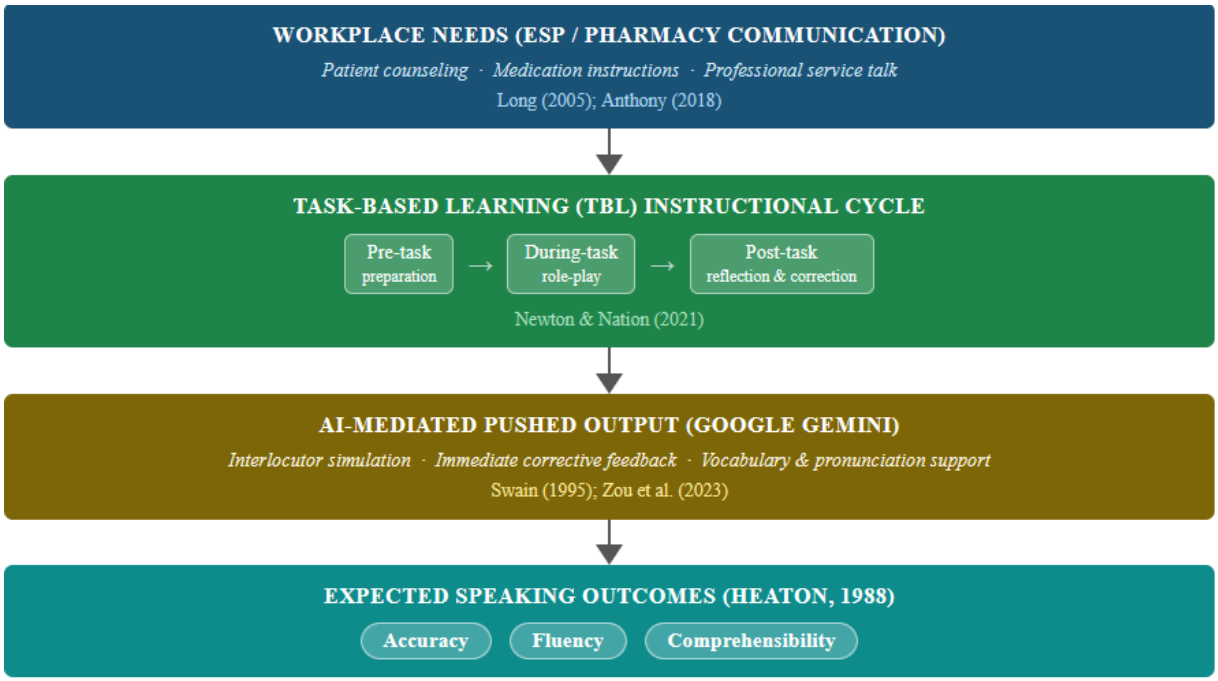
A complementary research on vocational contextualized learning integrates language instruction with authentic workplace tasks through systematic needs analysis (Long, 2005; Rahman et al., 2022). ESP supplies field-specific content that learners perceive as professionally meaningful (Anthony, 2018; Skarpaas & Hellekjær, 2021), while task-based learning (TBL) organizes speaking through pre-task preparation, during-task performance, and post-task reflection (Newton & Nation, 2021). Swain's (1995) Output Hypothesis adds that extended production promotes noticing and hypothesis testing, which may strengthen accuracy, fluency, and comprehensibility when learners revise their output. In SMK contexts, contextualized tasks appear to support engagement, yet teacher-led instruction often cannot provide immediate individual feedback at scale (Efrizah et al., 2024). ESP and TBL thus define what and how learners speak vocationally, while scalable corrective feedback, such as that enabled by AI tools reviewed above, may address a critical constraint in classroom-based speaking development.

Despite progress in AI-assisted and vocational English research, few studies examine their integration as a speaking intervention in Indonesian SMK settings. AI-speaking research continues to prioritise general proficiency in mainstream populations (Nguyen, 2024; Nhan et al., 2025), whereas vocational English research in Indonesia often emphasises needs analysis, curriculum design, or materials development over quasi-experimental speaking outcomes (Efrizah et al., 2024; Safira & Azzahra, 2022). Health-related vocational tracks, including pharmacy, remain underrepresented in experimental EFL literature (Anthony, 2018; Dudley-Evans & St. John, 1998). Synergistic models pairing accessible AI tools with workplace-relevant ESP tasks therefore lack robust empirical validation in SMK pharmacy classrooms.

The present study is grounded in an integrated conceptual framework synthesising ESP vocational contextualization, task-based learning, and AI-mediated pushed output (Figure 1). At the content level, ESP identifies pharmacy workplace needs as the basis for authentic speaking tasks (Long, 2005; Anthony, 2018). At the instructional level, TBL sequences pre-task, during-task, and post-task role-plays with feedback and output revision (Newton & Nation, 2021).

At the technological level, Google Gemini serves as interlocutor and corrective feedback provider, extending Swain's (1995) pushed-output mechanism through immediate post-production correction. Together, these elements predict improvement in accuracy, fluency, and comprehensibility (Heaton, 1988) when learners repeatedly perform vocationally authentic tasks with timely feedback. Pharmacy was selected as the focal field because patient counseling is a core workplace speaking demand (Anthony, 2018), SMKN 1 Polewali provided an intact Grade XI cohort with established ESP materials, and health-related SMK tracks remain underrepresented in AI-speaking research.

Figure 1. Conceptual Framework of AI-Based Vocational Speaking Instruction



Note. The framework illustrates how ESP pharmacy tasks, TBL sequencing, and AI-mediated pushed output converge to predict improvement in accuracy, fluency, and comprehensibility (Heaton, 1988).

To address the identified gap, this study investigated AI-based learning embedded in pharmacy vocational speaking tasks at SMKN 1 Polewali, guided by the following research questions:

1. Does AI-based learning in a vocational context significantly improve the speaking skills of Grade XI students at SMKN 1 Polewali in terms of accuracy, fluency, and comprehensibility?
2. Is there a significant difference in speaking skill gains between students taught through AI-based learning in a vocational context and students taught through conventional speaking instruction at SMKN 1 Polewali?
3. Which aspect of speaking skills shows the greatest improvement when using AI-based learning in a vocational context among Grade XI students at SMKN 1 Polewali?

Speaking performance was operationalized through Heaton's (1988) three-component rubric—accuracy, fluency, and comprehensibility. The novelty of this study lies in the synergistic integration of AI-based learning tools and vocational contextualized tasks, an approach that has not been empirically validated in the SMK pharmacy context in Indonesia. Findings are expected to inform ESP-oriented speaking instruction and the adaptation of AI-supported practice models for other SMK vocational fields.

2. Method

Research Design

A quantitative quasi-experimental approach was adopted to examine the effect of AI-based learning in a vocational pharmacy context on students' English speaking skills. Because the study involved intact classes that could not be randomly assigned, a nonequivalent pretest–posttest control group design was selected (Campbell & Stanley, 1966; Creswell & Creswell, 2018). This design permitted group comparison while documenting baseline proficiency through a speaking pretest and measuring post-

ISSN 2338-4778 (Print)
ISSN 2548-4192 (Online)

instruction change through an equivalent posttest (Shadish, Cook, & Campbell, 2002). Non-random assignment may introduce selection bias; however, comparable pretest overall means (experimental *M* = 36.94%; control *M* = 36.89%) supported baseline equivalence between groups.

The research was conducted at SMKN 1 Polewali, Polewali Mandar Regency, West Sulawesi, Indonesia, during the even semester of the 2025/2026 academic year (February–April 2026). Thirty-one active Grade XI Pharmacy students met the inclusion criteria: enrolment in the pharmacy programme, access to an internet-connected smartphone, and completion of foundational vocational coursework in Grade X. All participants were drawn from a single Grade XI Pharmacy class within the school. Because random assignment was not permitted within this intact homeroom structure, students were purposively assigned to an experimental subgroup (*n* = 15) and a control subgroup (*n* = 16) in consultation with the English teacher and school administration, with pretest scores used to verify baseline comparability. Purposive intact-class sampling was applied to maintain a single vocational field and ensure topical consistency for patient-counseling speaking tasks (Creswell & Creswell, 2018).

The experimental subgroup received Gemini-assisted vocational speaking instruction, whereas the control subgroup received conventional vocational-contextualised instruction without AI. The sample size (*N* = 31) limits statistical power and generalisability—a constraint typical of intact-classroom quasi-experiments—so findings should be interpreted with emphasis on effect sizes as well as *p*-values (Cohen, 1988). Nevertheless, the paired pre–post design increases sensitivity to within-group change, and large observed effects in the results suggest that substantive improvement was detected despite the modest sample.

The independent variable was AI-based learning in a vocational context, operationalised as the integration of Google Gemini with pharmacy-specific speaking tasks. During treatment, Gemini supported patient-counseling role-plays, corrective feedback, and vocabulary and pronunciation support aligned with pharmacy ESP communication. The dependent variable was speaking skill, measured using Heaton's (1988) rubric across accuracy (grammar, vocabulary, and pronunciation correctness), fluency (smooth pace with minimal unnatural pausing), and comprehensibility (intelligibility of the intended message).

Each component was scored on a 1–6 scale and converted into percentage scores interpreted using the national school-based classification (Kementerian Pendidikan dan Kebudayaan, 2017): 80–100% (Excellent), 69–79% (Good), 58–68% (Fair), 45–56% (Poor), and 35–44% (Very Poor). Speaking performance was assessed through a pharmacist–patient role-play task with follow-up questions in a patient-counseling context. The same test format was applied at pretest and posttest, while the scripts differed to reduce practice effects; the posttest used an over-the-counter medication scenario. Content validity of the task and rubric was reviewed by two English lecturers and one pharmacy teacher. Two English teachers rated all audio-recorded performances independently. Prior to scoring, both raters completed a half-day calibration workshop led by the researcher, during which they reviewed Heaton (1988) descriptor boundaries, examined sample performances, and independently scored five pilot recordings until discrepancies fell within one scoring band.

During the main scoring phase, raters worked in separate rooms without consultation. Inter-rater reliability for experimental-group posttest performances (*n* = 15) was examined using ICC(2,1), two-way mixed-effects, absolute agreement, single measures (Shrout & Fleiss, 1979). Agreement was moderate for accuracy (ICC = .725, 95% CI [.350, .899]), fluency (ICC = .601, 95% CI [.132, .847]), and comprehensibility (ICC = .671, 95% CI

[.248, .877]), with all $p \leq .009$ (Koo & Mae, 2016). Moderate ICC values are common in holistic speaking assessment; calibration and blind independent scoring were therefore implemented to reduce systematic rater bias.

Data collection proceeded in three phases. First, both subgroups completed a speaking pretest on a common assessment date; all performances were audio-recorded and rated independently by two raters. Second, treatment was delivered in eight 90-minute sessions over four weeks. Although both subgroups belonged to the same Grade XI Pharmacy class, they followed separate weekly English schedules during the intervention. This arrangement reduced the risk of cross-group contamination, especially unintended exposure of control students to Gemini-assisted activities.

In the experimental subgroup, students used Gemini on smartphones to complete five main speaking activities: a warm-up monologue, guided patient-counseling role-play, structured feedback requests, targeted vocabulary expansion, and a final unscripted role-play. The topics progressed from basic medication instructions to more complex counseling tasks, asking side effect, describe allergic including over-the-counter (OTC) consultation. In the control subgroup, students learned comparable pharmacy topics through conventional activities without AI, such as teacher modelling, pharmacist–patient videos, scripted dialogue practice, and unscripted role-plays. Both subgroups received the same number of sessions, duration, and syllabus-aligned topics.

Treatment fidelity in the experimental subgroup was monitored through a session checklist and field notes. The checklist was used to verify the completion of the five planned activity components, session duration, and topic coverage in each meeting. Field notes were used to document implementation quality, student responsiveness, and any adjustments made during the intervention. Field notes were also used to confirm comparable topic coverage in the control subgroup across the eight sessions. During the early treatment sessions, students had difficulty following Gemini-generated audio because the speech rate was too fast for their proficiency level. They also needed more time to understand the AI-generated dialogue texts. To address this issue, the teacher maintained the same instructional objectives and activity sequence but asked students to generate prompts at approximately A1 CEFR level and to request slightly slower audio output.

This adjustment made the dialogues easier to follow without changing the vocational focus of the treatment. In later meetings, particularly during the OTC consultation topic, another limitation emerged. Gemini's audio could not be replayed in short segments, so students could only replay the full dialogue rather than repeat it phrase by phrase. This reduced opportunities for detailed imitation and repetition practice. To manage this limitation, the teacher guided students through repeated whole-dialogue listening, note-taking, and oral rehearsal before performance. Because these adjustments did not change the instructional goals, the core task sequence, or the topical content, treatment fidelity was considered acceptable.

In the control subgroup, instruction covered comparable pharmacy topics through conventional activities without AI, including viewing pharmacist–patient videos, teacher modelling, pair work on scripted dialogues, and unscripted role-plays. Both subgroups received the same number of sessions, duration, and syllabus-aligned pharmacy topics. Treatment fidelity was monitored using a session checklist aligned with the experimental protocol, verifying completion of all five activity components and syllabus-aligned pharmacy topics in each session; field notes confirmed comparable topic coverage in the control subgroup across all eight sessions. Third, both subgroups completed a posttest on a common assessment date using the same format and scoring rubric as the pretest. Prior to data collection, institutional permission was administered from province from SMKN 1 Polewali. Students and guardians were informed of the study's purpose and their right to

ISSN 2338-4778 (Print)
ISSN 2548-4192 (Online)

participate voluntarily. Individual identities were kept confidential, and only group-level results were reported.

Data were analysed using IBM SPSS Statistics version 27 with a significance level of $\alpha = .05$. Normality of scores was assessed using Shapiro–Wilk and Kolmogorov–Smirnov tests. Raw pretest scores violated normality across all three aspects in both groups (Shapiro–Wilk $p < .001$ for all), consistent with a floor effect in which most students clustered in the Very Poor performance band at baseline. Composite posttest scores satisfied normality in both groups (experimental: $W = .927$, $p = .245$; control: $W = .956$, $p = .598$). Crucially, the composite gain score (posttest minus pretest) for the experimental group also met normality ($W = .895$, $p = .079$), directly satisfying the distributional assumption for paired-samples t-tests, which operate on within-subject difference scores rather than raw scores (Field, 2013; Pallant, 2020).

At the aspect level, gain scores showed mixed normality: fluency gain was normal ($W = .903$, $p = .106$), accuracy gain was borderline ($W = .882$, $p = .050$), and comprehensibility Data collection proceeded in three phases. First, both subgroups completed a speaking pretest on a common assessment date; all performances were audio-recorded and rated independently by two raters gain was non-normal ($W = .873$, $p = .038$).

Parametric tests were nonetheless retained for three reasons: the composite gain score driving the primary paired comparison met normality; paired t-tests remain robust to moderate violations with small, correlated repeated-measures samples (Field, 2013; Pallant, 2020); and the between-group independent-samples t-test was applied exclusively to composite posttest scores that satisfied normality. Because the between-group comparison relied solely on composite posttest scores, which met normality, the non-normal pretest distributions did not affect the validity of the independent-samples t-test. Levene's test confirmed homogeneity of variances across all score conditions, $F(5, 87) = 1.205$, $p = .314$, supporting use of the equal-variances assumed solution for the independent-samples t-test.

Paired-samples t-tests examined within-group pre–post changes for overall composite scores and each speaking aspect separately; an independent-samples t-test compared composite posttest performance between groups. Effect sizes were calculated using Cohen's d (Cohen, 1988; Sawilowsky, 2009). The speaking aspect with the greatest improvement in the experimental group was identified by comparing mean gain scores across accuracy, fluency, and comprehensibility.

3. Result

This section reports findings for three research questions concerning the effectiveness of AI-based learning in a vocational pharmacy context (RQ1), comparative post-test performance between groups (RQ2), and the aspect showing the greatest improvement in the experimental group (RQ3). Results are presented in the order of assumption testing and inferential analyses specified in the Method section.

Classification of Speaking Performance

Frequency distributions indicated that both groups began the study with predominantly Very Poor speaking classifications across aspects. At pretest, 86.7% of experimental students and 81.3% of control students were classified as Very Poor in accuracy. Following intervention, the experimental group shifted toward Fair and Good post-test categories across aspects, whereas the control group showed a more modest shift, remaining largely in Poor or Fair categories at posttest. These patterns are consistent with comparable baseline entry and divergent post-treatment trajectories.

Descriptive Statistics

Table 1 presents pretest and posttest means, standard deviations, gains, and post-test classifications by group and aspect.

Table 1. Descriptive Statistics for Speaking Scores by Group and Aspect

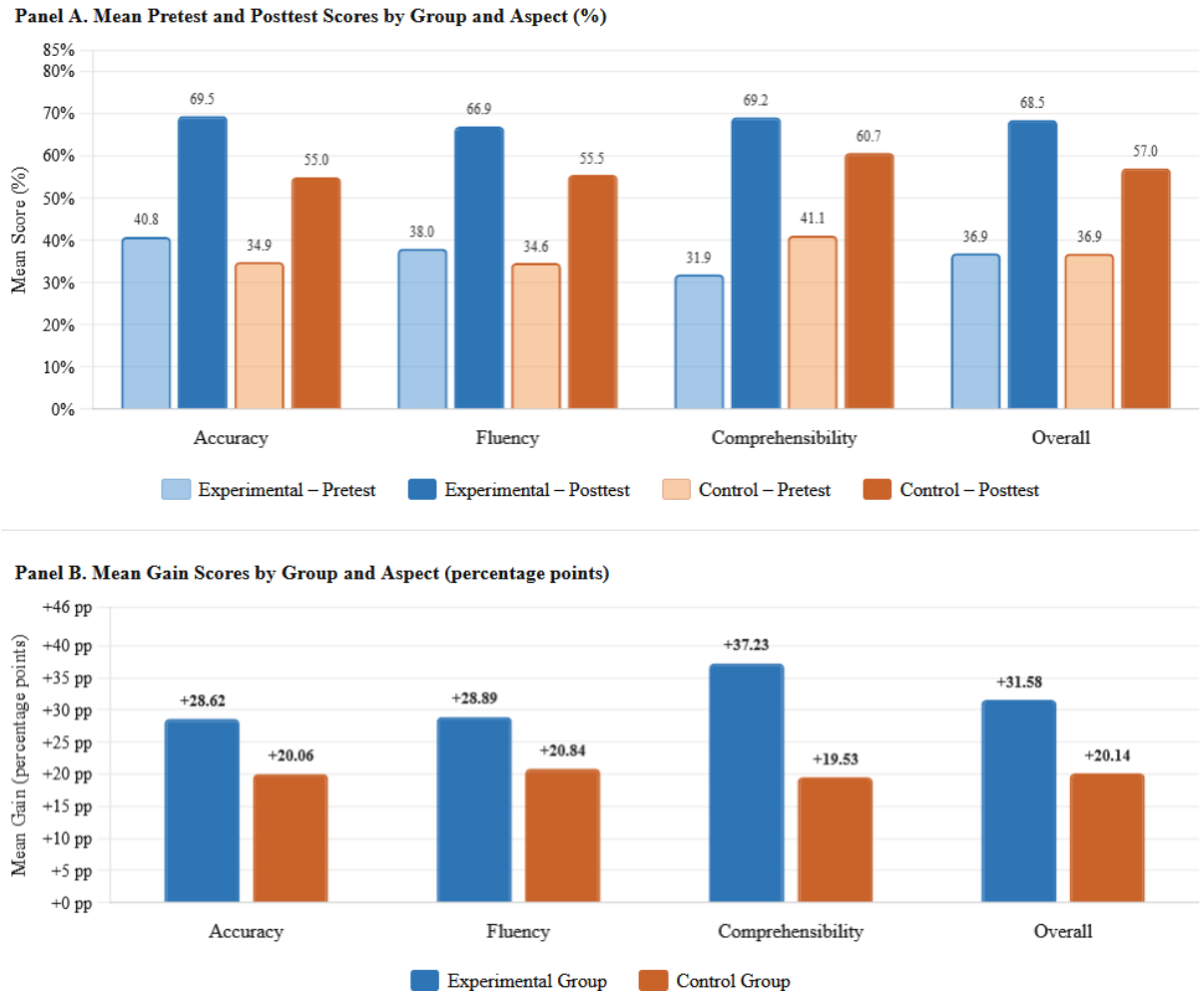
Group / Aspect	Pretest *M*	Pretest SD	Posttest *M*	Posttest SD	Gain (pp)	Post-test class.[a]
Experimental (*n* = 15)						
Accuracy	40.83	12.62	69.45	15.32	+28.62	Good
Fluency	38.05	12.29	66.94	11.30	+28.89	Fair
Comprehensibility	31.94	11.96	69.17	12.68	+37.23	Good
Overall	36.94	—	68.52	11.96	+31.58	Fair
Control (*n* = 16)						
Accuracy	34.89	11.87	54.95	10.89	+20.06	Poor
Fluency	34.63	8.57	55.47	13.15	+20.84	Poor
Comprehensibility	41.15	7.74	60.68	8.60	+19.53	Fair
Overall	36.89	—	57.03	9.62	+20.14	Poor

Note. M = mean percentage score; SD = standard deviation; Gain = posttest *M* minus pretest *M*; pp = percentage points. A Classification per Kementerian Pendidikan dan Kebudayaan (2017): 80–100% Excellent; 69–79% Good; 58–68% Fair; 45–56% Poor; 35–44% Very Poor.

As shown in Table 1, both groups entered the study at comparable overall proficiency (experimental *M* = 36.94%; control *M* = 36.89%). After treatment, the experimental group attained an overall posttest mean of 68.52% (Fair), whereas the control group reached 57.03% (Poor). The experimental group recorded the largest absolute gain in comprehensibility (+37.23 pp), followed by fluency (+28.89 pp) and accuracy (+28.62 pp). At posttest, the widest between-group gap by aspect was accuracy (+14.50 pp favouring the experimental group).

Figure 3 provides a visual representation of these patterns. Panel A displays pretest and posttest mean scores for both groups across all three aspects and the overall composite score, illustrating divergent post-treatment trajectories despite comparable baseline entry levels. Panel B presents mean gain scores by group and aspect, visually confirming that the experimental group achieved substantially larger gains on all aspects, with the comprehensibility aspect showing the greatest improvement (+37.23 pp vs. +19.53 pp for the control group).

ISSN 2338-4778 (Print)
ISSN 2548-4192 (Online)



Note. Values are mean percentage scores (%) based on Heaton's (1988) three-component speaking rubric (accuracy, fluency, comprehensibility) scored on a 1–6 scale and converted to percentages. Classification thresholds (Kementerian Pendidikan dan Kebudayaan, 2017): 80–100% Excellent; 69–79% Good; 58–68% Fair; 45–56% Poor; 35–44% Very Poor. Gain = posttest mean − pretest mean (percentage points). Exp = Experimental group (n = 15); Con = Control group (n = 16). Dashed horizontal lines in Panel A indicate classification band boundaries (44%, 56%, 68%, 79%).

Inter-Rater Reliability

Inter-rater reliability for experimental-group posttest scores is reported in Table 2.

Table 2. Inter-Rater Reliability (ICC) for Experimental Post-Test Scores (n = 15)

Aspect	ICC(2,1)	95% CI	*F*	*p*	Interpretation
Accuracy	.725	[.350, .899]	5.966	<.001	Moderate
Fluency	.601	[.132, .847]	3.822	.009	Moderate
Comprehensibility	.671	[.248, .877]	4.812	.003	Moderate
Overall mean	.666	—	—	—	Moderate

Note. ICC = intraclass correlation coefficient; two-way mixed effects model, absolute agreement, single measures (Shrout & Fleiss, 1979; Koo & Mae, 2016).

These moderate ICC values (mean ICC = .666) confirm that both raters assigned comparable scores across performances, with all reliability tests reaching significance ($p \leq .009$), indicating consistent rater discrimination across proficiency levels. In practical terms, the level of agreement achieved is sufficient for group-level score interpretation in holistic EFL speaking assessment, particularly given that calibration procedures were implemented prior to main-phase scoring to align rater judgements.

Normality and Homogeneity of Variance

Normality of composite overall speaking scores was examined using the Kolmogorov–Smirnov (K–S) and Shapiro–Wilk (S–W) tests (Table 3). Pretest distributions for both groups violated normality (all S–W $p < .001$). Posttest distributions met normality (experimental S–W $p = .245$; control S–W $p = .598$). Given small sample sizes and the robustness of paired comparisons for correlated pre–post data, parametric tests were retained (Field, 2013; Pallant, 2020).

Table 3. Normality Test Results for Overall Speaking Scores

Group	K–S stat	K–S *p*	S–W stat	S–W *p*	Decision
Pretest (Exp.)	.255	.010	.656	<.001	Not normal
Posttest (Exp.)	.197	.120	.927	.245	Normal
Pretest (Con.)	.349	<.001	.678	<.001	Not normal
Posttest (Con.)	.196	.100	.956	.598	Normal

Note. S–W = Shapiro–Wilk; K–S = Kolmogorov–Smirnov with Lilliefors correction. α = .05.

The non-normal pretest distributions reflect a floor effect in which most students entered the study with limited exposure to pharmacy-contextualized English speaking tasks, resulting in scores concentrated in the Very Poor band. This clustering confirms genuine baseline comparability between groups rather than a measurement artifact. The shift to normal posttest distributions indicates that instruction produced meaningful differentiation across the performance range in both groups.

Levene's test indicated homogeneity of variances for overall scores, $F(5, 87) = 1.205$, $p = 0.314$ (Table 4). The equal variances assumed solution was therefore used for the independent samples *t*-test.

Table 4. Levene's Test for Equality of Variances (Overall Speaking Scores)

Basis	Levene *F*	df1	df2	*p*
Based on mean	1.205	5	87	0.314
Based on median	.984	5	87	0.432
Based on trimmed mean	1.154	5	87	.338

The non-significant Levene's test ($p = .314$) confirmed variance homogeneity across groups, supporting the validity of subsequent between-group comparisons.

Paired-Samples t-Tests (RQ1)

Table 5 reports within-group pre–post comparisons for the experimental group. Significant improvements were found for accuracy, $t(14) = -9.106$, $p < .001$, mean difference = −28.80, 95% CI [−35.58, −22.02]; fluency, $t(14) = -12.860$, $p < .001$, mean

ISSN 2338-4778 (Print)
ISSN 2548-4192 (Online)

difference = −29.13, 95% CI [−33.99, −24.28]; and comprehensibility, $t(14) = -12.506$, $p < .001$, mean difference = −37.33, 95% CI [−43.74, −30.93]. Negative mean differences reflect higher posttest scores in SPSS pre-minus-post coding.

These results answer RQ1 affirmatively: AI-based learning in a vocational context significantly improved speaking skills across all three measured aspects in the experimental group. Practically, the overall mean gain of 31.58 percentage points represents an advancement of two full classification bands — from Very Poor (pre M = 36.94%) to Fair (post M = 68.52%) — indicating educationally meaningful change that extends beyond statistical significance. Within-group effect sizes were huge for all three aspects (accuracy d = 2.35, fluency d = 3.32, comprehensibility d = 3.23; average d = 2.97; Sawilowsky, 2009), confirming that improvements were of substantial magnitude relative to within-group variability.

Table 5. Paired-Samples *t*-Test — Experimental Group (n = 15)

Pair	Comparison	*M* diff	SD	SE	95% CI	*t*	df	*p*
1	Accuracy (Pre–Post)	−28.800	12.249	3.163	[−35.58, −22.02]	−9.106	14	<.001
2	Fluency (Pre–Post)	−29.133	8.774	2.265	[−33.99, −24.28]	−12.860	14	<.001
3	Comprehensibility	−37.333	11.561	2.985	[−43.74, −30.93]	−12.506	14	<.001

The control group also improved significantly on all aspects (Table 6): accuracy, $t(15) = -6.717$, $p < .001$; fluency, $t(15) = -7.248$, $p < .001$; comprehensibility, $t(15) = -7.324$, $p < .001$. However, experimental gains exceeded control gains on every aspect (accuracy: +28.80 vs +20.06 pp; fluency: +28.89 vs +20.84 pp; comprehensibility: +37.33 vs +19.53 pp).

Table 6. Paired-Samples *t*-Test — Control Group (n = 16)

Pair	Comparison	*M* diff	SD	SE	95% CI	*t*	df	*p*
4	Accuracy (Pre–Post)	−20.063	11.947	2.987	[−26.43, −13.70]	−6.717	15	<.001
5	Fluency (Pre–Post)	−21.063	11.625	2.906	[−27.26, −14.87]	−7.248	15	<.001
6	Comprehensibility	−19.375	10.582	2.646	[−25.01, −13.74]	−7.324	15	<.001

Although the control group's within-group gains also reflected very large practical effects (average d = 1.77), posttest means for accuracy (54.95%) and fluency (55.47%) remained in the Poor classification — one full band below the experimental group's corresponding Good and Fair posttest attainment. This differential indicates that conventional vocational instruction, while statistically effective, did not consistently elevate students to higher performance categories within the study period.

Independent-Samples t-Test (RQ2)

1. Independent-Samples t-Test (RQ2)

Table 7 presents the independent-samples *t*-test results comparing overall posttest speaking scores between the experimental and control groups. The experimental group (M = 68.52%, SD = 11.96) significantly outperformed the control group (M = 57.03%, SD = 9.62), $t(29) = 2.95$, $p = .006$ (two-tailed), mean difference = 11.49 pp, 95% CI [3.54, 19.44].

These results answer RQ2: students who received AI-based vocational instruction achieved significantly higher posttest speaking performance than those who received conventional instruction.

Beyond statistical significance, this 11.49-percentage-point difference corresponds to a large between-group effect size (d = 1.06; Cohen, 1988), representing a shift of one full classification band between groups. In practical terms, the average experimental student at posttest reached the Fair band (68.52%), while the average control student remained in the Poor band (57.03%) — a classroom-observable performance advantage associated with AI-based instruction.

Table 7. Independent-Samples *t*-Test: Post-Test Overall Speaking Scores (Experimental vs. Control)

Group	N	*M* (%)	SD	SE	*t*	df	*p*	M diff (pp)	95% CI
Experimental (AI-based)	15	68.52	11.96	3.09	2.95	29	**.006**	11.49	[3.54, 19.44]
Control (Conventional)	16	57.03	9.62	2.41	—	—	—	—	—

Note. Equal Variances Assumed; Levene's test confirmed variance homogeneity (F = 1.205, p = .314; see Table 4). *M* = Mean; SD = Standard Deviation; SE = Standard Error; M diff = Mean difference (Experimental − Control); pp = percentage points; CI = Confidence Interval. Source: SPSS Output.

Effect Sizes (Cohen's d)

Within-group effect sizes for the experimental group were huge for all aspects: accuracy d = 2.35, fluency d = 3.32,

An independent-samples *t*-test on overall posttest scores showed that the experimental group (*M* = 68.52%, SD = 11.96) significantly outperformed the control group (*M* = 57.03%, SD = 9.62), $t(29) = 2.95$, $p = .006$ (two-tailed), mean difference = 11.49 pp, 95% CI [3.54, 19.44]. This result answers RQ2: students who received AI-based vocational instruction achieved significantly higher post-test speaking performance than those who received conventional instruction.

Effect Sizes (Cohen's d)

Within-group effect sizes for the experimental group were huge for all aspects: accuracy d = 2.35, fluency d = 3.32, comprehensibility d = 3.23, with an average d = 2.97 (Table 7). Control-group within-group effects were very large (average d = 1.77). Between-group posttest effect sizes favoured the experimental group: overall d = 1.06 (large), accuracy d = 1.10 (large), fluency d = 0.93 (large), and comprehensibility d = 0.79 (medium–large).

Table 7. Cohen's *d* Effect Sizes

Comparison	Aspect	Statistic	Cohen's *d*	Category
Within-group (experimental, *n* = 15)	Accuracy	$\|t\| = 9.106$	2.35	Huge
	Fluency	$\|t\| = 12.860$	3.32	Huge
	Comprehensibility	$\|t\| = 12.506$	3.23	Huge

ISSN 2338-4778 (Print)
ISSN 2548-4192 (Online)

Comparison	Aspect	Statistic	Cohen's *d*	Category
Within-group (control, *n* = 16)	Average	—	2.97	Huge
	Accuracy	$\|t\| = 6.717$	1.68	Very large
	Fluency	$\|t\| = 7.248$	1.81	Very large
	Comprehensibility	$\|t\| = 7.324$	1.83	Very large
Between-group (posttest)	Average	—	1.77	Very large
	Accuracy	*M* 69.45% vs 54.95%	1.10	Large
	Fluency	*M* 66.94% vs 55.47%	0.93	Large
	Comprehensibility	*M* 69.17% vs 60.68%	0.79	Medium–large
	Overall	*M* 68.52% vs 57.03%	1.06	Large

Note. Within-group: $d = |t| / \sqrt{n}$ (Cohen, 1988). Between-group: pooled SD. Categories: Sawilowsky (2009) for within-group huge ($d \geq 2.00$); Cohen (1988) for between-group large ($d \geq 0.80$).

Greatest Improved Aspect (RQ3)

Descriptive gain scores in the experimental group indicate that comprehensibility showed the greatest improvement (+37.23 pp), followed by fluency (+28.89 pp) and accuracy (+28.62 pp). This ordering answers RQ3 directly. The comprehensibility finding is further supported by the largest within-group effect size for that aspect in absolute terms relative to its paired *t* magnitude ($d = 3.23$, huge), although fluency recorded the highest *d* value (3.32).

Summary of Findings

In summary, AI-based vocational speaking instruction produced significant and large within-group gains (RQ1), significantly higher posttest performance than conventional instruction (RQ2), and the largest absolute improvement in comprehensibility (RQ3). Conventional instruction also yielded significant within-group progress, confirming the value of vocational contextualization, but with smaller gains and lower posttest attainment than the AI-supported condition.

4. Discussion

The findings of this study consistently support the effectiveness of AI-based learning in a vocational context for improving Grade XI Pharmacy students' English speaking performance. All three research questions are answered affirmatively: the experimental group achieved significant and practically large pre-post gains on every measured aspect of speaking (RQ1), significantly outperformed the control group at posttest (RQ2), and recorded the greatest absolute improvement in comprehensibility among the three aspects (RQ3). The control group also improved significantly, confirming that vocational

contextualization itself constitutes a viable instructional condition for speaking development. The comparative advantage of AI-assisted instruction, however, was evident across all aspects and particularly pronounced in comprehensibility as an outcome that invites theoretical interpretation.

These results may be interpreted through Swain's (1995) Output Hypothesis, which treats oral production as a driver of language learning rather than merely an outcome. Pharmacy patient-counseling role-plays required learners to produce pushed output: extended explanations of dosage, side effects, drug types, allergic reactions, and OTC recommendations under communicative pressure. Gemini functioned as an interlocutor and feedback provider that made gaps between intended and actual performance visible at the moment of production. The structured error-review phase is consistent with the noticing and hypothesis-testing processes described by Swain (1995): students systematically revised grammar, vocabulary, and phrasing after each role-play.

Although these cognitive mechanisms were inferred from the instructional design rather than directly observed or measured. The disproportionate comprehensibility gain is particularly consistent with this framework, because pharmacy service communication depends on whether the listener can extract the speaker's intended message without repeated clarification. When learners were prompted to deliver complete, intelligible counseling utterances and received immediate feedback on clarity, they appeared to allocate greater attention to message packaging, not only formal accuracy.

The instructional design also aligns with task-based learning principles articulated by Newton and Nation (2021). Eight vocational scenarios supplied familiar, field-specific content that became increasingly complex from medication instructions to multi-medication counseling. Repeated performance within and across sessions supported fluency development through time-on-task and meaningful practice rather than isolated drills. The warm-up monologue phase encouraged message-focused output without immediate correction, whereas the structured error-review phase resembled TBL reflection in which form-focused work emerges after communicative performance. AI-mediated interaction may also have extended practice beyond teacher-fronted classroom time; several studies suggest that chatbot interaction tends to reduce evaluative pressure relative to peer-present classroom performance (Nhan et al., 2025; Nguyen, 2024), though this mechanism was not directly measured in the present study.

Comparison with prior studies strengthens confidence in the pattern while clarifying scope. The significant within-group improvements and large effect sizes align with the systematic review by Safitri et al. (2025), who synthesised 18 studies and found consistent evidence that AI chatbots enhance speaking proficiency through conversational simulations, instant corrective feedback, and personalised practice, mechanisms directly mirrored by Gemini's role in the present intervention. Safitri et al. (2025) further document that chatbots support fluency through reduced speech pauses and pronunciation improvement through real-time feedback, patterns consistent with the gains observed here. Nhan et al. (2025) similarly propose that reduced anxiety and personalised practice may promote speaking participation and sustained engagement in AI-assisted settings; although speaking anxiety and student engagement were not measured in the present study, these mechanisms offer a plausible theoretical account for the observed pattern of improvement and warrant direct empirical investigation in future research. Nguyen (2024) cautions, however, that long-term retention of AI-mediated gains remains understudied — a limitation that also applies to the present four-week intervention. Chen et al. (2023) offer an important methodological contrast: AI scoring of fluency may diverge from human holistic judgement.

ISSN 2338-4778 (Print)
ISSN 2548-4192 (Online)

The present study therefore retained trained human raters using Heaton's (1988) rubric, with moderate but statistically significant inter-rater agreement (ICC .60–.73), strengthening the validity of reported gains as human-judged communicative performance rather than machine-generated scores.

Theoretically, the study extends AI-speaking research into Indonesian ESP pharmacy education by validating, for the first time in this context, the synergistic integration of accessible AI tools and vocational contextualized tasks. The novelty of this study lies in the synergistic integration of AI-based learning tools and vocational contextualized tasks as an approach that has not been empirically validated in the SMK pharmacy context in Indonesia. Practically, the findings suggest that Gemini and comparable chatbots can serve as supplements, not replacements for teacher-led instruction by supplying immediate feedback and additional rehearsal on workplace-relevant scripts. Teachers can structure AI-assisted speaking sessions around the four-phase format demonstrated in this study: warm-up monologue, guided AI role-play, structured error review, and vocabulary enhancement, to situate AI feedback within a pedagogically coherent vocational task sequence.

These practical implications highlight a pressing need for deliberate teacher professional development (PD). Effective AI integration in vocational speaking instruction requires teachers to develop competencies that extend beyond conventional English language pedagogy. Specifically, three professional capacities warrant systematic development: (a) task-design literacy — the ability to design and sequence AI-compatible speaking tasks aligned with vocational ESP needs, so that AI interaction remains grounded in authentic workplace communication rather than decontextualised exchange; (b) prompt-engineering literacy — the ability to guide students in constructing effective prompts, adjusting AI output complexity to match learner proficiency (as the A1 CEFR adjustment in this study illustrates), and managing AI interaction quality; and (c) feedback evaluation capacity — the critical ability to review AI-generated corrections for accuracy and professional register appropriateness before students internalise them, given that large language models may produce plausible but incorrect output.

Without structured PD in these areas, teachers risk either uncritically delegating corrective authority to AI tools or avoiding them altogether due to uncertainty about their pedagogical value. School leaders and curriculum designers at SMK institutions are encouraged to embed AI-literacy components into in-service teacher training, develop peer-learning communities centred on AI-integrated speaking instruction, and allocate dedicated time for teachers to pilot, reflect on, and iteratively refine AI-assisted lesson designs. At policy level, the results are broadly consistent with Permendikdasmen Number 13 of 2025, which encourages AI integration in vocational education to strengthen industry-relevant competencies; sustainable and equitable implementation, however, requires that national and regional education authorities invest not only in infrastructure and device access, but also in systematic teacher PD frameworks that address the pedagogical, technical, and ethical dimensions of AI use in vocational English classrooms.

Notwithstanding the positive outcomes, several critical considerations surrounding AI integration in EFL speaking instruction warrant explicit discussion. First, sustained use of AI interlocutors carries the risk of learner over-reliance: students who routinely practice with a highly patient, non-evaluative chatbot may develop speaking strategies calibrated to AI interaction patterns, which may not transfer readily to the variable demands of authentic human conversation (Lee, 2023; Nhan et al., 2025). The absence of authentic social interaction dynamics, including turn-taking pressure and negotiation of meaning in chatbot interaction may leave certain real-world speaking competencies underdeveloped (Fathi et

al., 2024; Nhan et al., 2025). Second, AI-generated feedback including that produced by Gemini is not inherently accurate. Large language models may produce plausible but grammatically or lexically inappropriate corrections, a risk particularly consequential in professional register development where precision matters (Chen et al., 2023; Nguyen, 2024).

In the present study, no systematic verification of Gemini's corrective feedback quality was conducted, which represents a limitation that future studies should address through instructor review of AI feedback logs. Third, practical platform constraints were directly observed during implementation. In early sessions, Gemini's audio output was generated at a speech rate beyond the learners' proficiency level, requiring the teacher to instruct students to generate prompts at approximately A1 CEFR level and request slower audio output. In later sessions, the inability to replay audio in short segments — only full-dialogue replay was available — reduced opportunities for detailed imitation and phrase-by-phrase repetition practice; the teacher compensated through whole-dialogue listening, note-taking, and oral rehearsal.

These constraints, while managed without altering core instructional objectives, underscore that AI tool usability in resource-constrained SMK settings is not automatic and requires deliberate pedagogical scaffolding (Arora et al., 2024; Pishadast, 2022). Fourth, digital literacy and device access remain unequally distributed across Indonesian SMK settings (Nuryadin & Marlina, 2023). Students unfamiliar with prompt engineering or chatbot interaction conventions may allocate cognitive resources to platform navigation rather than speaking practice itself (Sweller, 1994). Teachers implementing AI-assisted speaking instruction should therefore incorporate explicit platform orientation and monitor the quality of student–AI interaction, not merely speaking output scores.

Several limitations warrant cautious generalization. The sample was small (n = 15 and n = 16) and drawn from a single institution using intact classes without random assignment, although pretest equivalence supported baseline comparability. The treatment lasted only four weeks, so long-term retention of speaking gains remains unknown. Only one AI platform (Gemini) was examined; moreover, two platform-specific constraints emerged during implementation audio output at a speech rate exceeding learner proficiency and the inability to replay audio in short segments, that required mid-course pedagogical adjustments and may have limited certain imitation and repetition opportunities.

The quality and accuracy of Gemini's corrective feedback were not independently verified, which limits mechanistic conclusions about the source of improvement. Moderate inter-rater reliability, especially for fluency (ICC = .601), reflects the inherent subjectivity of holistic speaking assessment. Finally, affective variables such as speaking anxiety and student engagement were not measured in this quantitative design; claims about these mechanisms remain inferential and should be tested empirically in future research.

Future research should replicate the model with larger, multi-site samples across SMK regions and vocational fields such as nursing, hospitality, and engineering technology. Longitudinal designs with delayed posttests would clarify retention of speaking gains after treatment ends. Comparative studies of multiple AI tools could identify which feedback features most strongly support accuracy, fluency, and comprehensibility in vocational ESP. Systematic analysis of AI feedback logs would establish whether chatbot-generated corrections are consistently accurate and professionally appropriate for pharmacy register. Mixed-methods investigations are recommended to document learners' affective experiences, digital literacy challenges, and teachers' implementation decisions in resource-constrained settings. Such work would strengthen the evidence base for scaling AI-supported vocational speaking instruction while directly addressing the dependency, feedback accuracy, and equity concerns identified above.

ISSN 2338-4778 (Print)
ISSN 2548-4192 (Online)

5. Conclusion

This study examined whether AI-based learning embedded in a vocational pharmacy context could improve the English speaking skills of Grade XI students at SMKN 1 Polewali. The findings indicate that Gemini-assisted vocational speaking instruction was associated with significant improvement in students' speaking performance across accuracy, fluency, and comprehensibility (RQ1). In addition, students in the AI-supported condition achieved higher post-instruction speaking outcomes than those in the conventional condition (RQ2). Comprehensibility showed the greatest improvement among the assessed aspects (RQ3), suggesting that AI-mediated role-play may be particularly supportive for clearer workplace-oriented communication in patient-counseling tasks. The novelty of this study lies in the synergistic integration of AI-based learning tools and vocational contextualized tasks, an approach that has not been empirically validated in the SMK pharmacy context in Indonesia.

This conclusion should be interpreted with caution. The findings are limited by the small sample size, the use of intact classes in a single school setting, and the short intervention duration. These constraints may restrict the generalisability of the results beyond similar SMK pharmacy contexts.

Suggestions and recommendations

For classroom practice, teachers are encouraged to use accessible AI chatbots as supplementary tools to extend speaking rehearsal and provide timely feedback, while keeping instruction anchored in authentic vocational scenarios such as patient counseling. Curriculum designers in SMK contexts are advised to prioritise workplace-specific speaking tasks over generic conversation topics to strengthen the relevance of speaking instruction to industry communication demands. School leaders and policymakers may align implementation with national directions on AI integration in vocational education, while also supporting teacher development, connectivity, and ethical guidelines to promote equitable and sustainable adoption.

For further research, replication studies with larger and multi-site samples across different SMK vocational programmes are recommended to strengthen external validity. Longitudinal designs with delayed posttests are needed to examine retention of speaking gains after AI-supported instruction. Future studies may also compare multiple AI tools and incorporate qualitative or mixed-methods data to document learners' experiences and implementation challenges in low-resource settings.

6. Acknowledgement

The authors express sincere gratitude to the principal, English teachers, and Grade XI Pharmacy students of SMKN 1 Polewali for their cooperation and participation in this study. Appreciation is also extended to the Postgraduate Program of English Language Education, State University of Makassar (UNM), for academic support during the research process. Any remaining errors are the responsibility of the authors.

7. References

Anthony, L. (2018). *Introducing English for specific purposes*. Routledge.

Arora, B., Al-Wadi, H., & Afari, E. (2024). Scaffolding instruction for improvement in learning English language skills. *International Journal of Evaluation and Research in Education (IJERE), 13*(2), 1265–1275. doi:10.11591/ijere.v13i2.26659

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research*. Rand McNally.

Chen, J., Lai, P., Chan, A., Man, V., & Chan, C.-H. (2023). AI-assisted enhancement of student presentation skills: Challenges and opportunities. *Sustainability, 15*(1), Article 196. doi:10.3390/su15010196

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.). SAGE.

Dudley-Evans, T., & St. John, M. J. (1998). *Developments in English for specific purposes: A multi-disciplinary approach*. Cambridge University Press.

Efrizah, D., Fadly, Y., & Putri, V. O. (2024). English speaking barriers in vocational education: A study of SMK SPP SNAKMA students. *Journal of Language, 6*(2), 271–278. e-ISSN: 2685-8878 | p-ISSN: 2655-9080. Retrieved from https://jurnal.uisu.ac.id/index.php/journaloflanguage

Fathi, J., Rahimi, M., & Derakshan, A. (2024). Improving EFL learners' speaking skills and willingness to communicate via artificial intelligence-mediated interactions. *System, 121*, Article 103254. doi:10.1016/j.system.2024.103254

Field, A. (2013). *Discovering statistics using IBM SPSS Statistics* (4th ed.). SAGE.

Heaton, J. B. (1988). *Writing English language tests*. Longman.

Kementerian Pendidikan dan Kebudayaan. (2017). Peraturan Menteri Pendidikan dan Kebudayaan Republik Indonesia nomor 3 tahun 2017 tentang penilaian hasil belajar oleh pemerintah dan penilaian hasil belajar oleh satuan pendidikan. *Berita Negara Republik Indonesia Tahun 2017 nomor 117*.

Kementerian Pendidikan Dasar dan Menengah. (2025). Peraturan Menteri Pendidikan Dasar dan Menengah Republik Indonesia Nomor 13 Tahun 2025 tentang Perubahan atas Peraturan Menteri Pendidikan, Kebudayaan, Riset, dan Teknologi Nomor 12 Tahun 2024 tentang Kurikulum pada Pendidikan Anak Usia Dini, Jenjang Pendidikan Dasar, dan Jenjang Pendidikan Menengah. *Berita Negara Republik Indonesia Tahun 2025 Nomor 503*. Retrieved from https://jdih.kemendikdasmen.go.id

Koo, T. K., & Mae, A. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155–163. doi:10.1016/j.jcm.2016.02.012

Lee, S.-E. (2023). Otherwise than teaching by artificial intelligence. *Journal of Philosophy of Education, 57*, 553–570. https://doi.org/10.1093/jopedu/qhad019

Lima, L. A. de O., Gomes, L. P., Silva e Silva, P. H. da, Oliveira, E. F. da S., Nascimento, M. do, Tourem, R. V., Gonçalves, J. N. de A., Lima, A. da S., Sobral, R., & Santos, I. da M. P. dos. (2024). Artificial intelligence and its use in the educational process. *Navigating Through the Knowledge of Education, 43*, 1–20. doi:10.56238/sevened2024.002-043

Long, M. H. (Ed.). (2005). *Second language needs analysis*. Cambridge University Press.

Newton, J. M., & Nation, I. S. P. (2021). *Teaching ESL/EFL listening and speaking* (2nd ed.). Routledge.

Nguyen, H. A. (2024). Harnessing AI-based tools for enhancing English speaking proficiency: Impacts, challenges, and long-term engagement. *International Journal of AI in Language Education, 1*(2), 18–29. doi:10.54855/ijaile.24122

Nhan, L. K., Hoa, N. T. M., & Quang, L. V. N. (2025). Revolutionizing speaking skills improvement: AI's role in personalized language learning. *International Journal of Innovative Research and Scientific Studies, 8*(2), 4637–4649. doi:10.53894/ijirss.v8i2.6408

Nuryadin, R., & Marlina. (2023). The use of artificial intelligence in education (literature review). *Indonesian Journal of Primary Education, 7*(2), 143–156. Retrieved from https://ejournal.upi.edu/index.php/IJPE/index

Pallant, J. (2020). *SPSS survival manual: A step by step guide to data analysis using IBM SPSS* (7th ed.). Routledge.

ISSN 2338-4778 (Print)
ISSN 2548-4192 (Online)

Pishadast, A. (2022). Developing the speaking ability of EFL learners through scaffolding. *Journal of Contemporary Language Research (JCLR), 1*(2), 60–64. doi:10.58803/jclr.v1i2.8

Rahman, S., Yani, A., & Prasetyo, Z. K. (2022). Vocational English contextualization in Industry 4.0. *Journal of Vocational Education*, *18*(2), 90–104.

Safira, L., & Azzahra, N. F. (2022). Meningkatkan kesiapan kerja lulusan SMK melalui perbaikan kurikulum bahasa Inggris. Center for Indonesian Policy Studies. doi:10.35497/558654

Safitri, E. I., Hidayati, S., & Ciptaningrum, D. S. (2025). The impact of AI chatbots on English language learners' speaking proficiency: A systematic review. *Journal of Research on English and Language Learning (J-REaLL)*, *6*(2), 317–329. https://doi.org/10.33474/j-reall.v6i2.23866

Sawilowsky, S. S. (2009). New effect size rules of thumb. *Journal of Modern Applied Statistical Methods, 8*(2), 597–599. doi:10.22237/jmasm/1257035100

Seidlhofer, B. (2011). *Understanding English as a lingua franca*. Oxford University Press.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420–428. doi:10.1037/0033-2909.86.2.420

Skarpaas, K. G., & Hellekjær, G. O. (2021). Vocational orientation—A supportive approach to teaching L2 English in upper secondary school vocational programmes. *International Journal of Educational Research Open, 2*, Article 100064. doi:10.1016/j.ijedro.2021.100064

Statistics Indonesia. (2021). Unemployment rate by education level, 2019–2021. Retrieved from https://www.bps.go.id/indicator/6/1179/1/unemployment-rate-by-education-level.html

Swain, M. (1995). Three functions of output in second language learning. In G. Cook & B. Seidlhofer (Eds.), *Principle and practice in applied linguistics: Studies in honour of H. G. Widdowson* (pp. 125–144). Oxford University Press.

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction, 4*(4), 295–312. doi:10.1016/0959-4752(94)90003-5

Thompson, F., & Lim, G. (2021). *Reaping the benefits of industry 4.0 through skills development in high-growth industries in Southeast Asia: Insights from Cambodia, Indonesia, the Philippines, and Viet Nam* (Report No. SPR-200328). Asian Development Bank. doi:10.22617/SPR200328

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education—Where are the educators? *International Journal of Educational Technology in Higher Education, 16*, Article 39. doi:10.1186/s41239-019-0171-0

Zou, B., Du, Y., Wang, Z., Chen, J., & Zhang, W. (2023). An investigation into artificial intelligence speech evaluation programs with automatic feedback for developing EFL learners' speaking skills. *SAGE Open, 13*(3), 1–9. doi:10.1177/21582440231193818